Within lake clustering of high resolution satellite retrievals - a functional data and clustering approach

R. O'Donnell[1], C. Miller[1], E.M. Scott[1]

[1] University of Glasgow, UK

E-mail for correspondence: ruth.haggarty@glasgow.ac.uk

Abstract: Satellite retrievals for lakes provide high resolution spatiotemporal images. This paper proposes a within lake dimensionality reduction approach using functional data analysis to enable computationally efficient clustering of temporal patterns within lakes.

Keywords: Remote sensing; dimensionality reduction; functional data; clustering.

1 Introduction

Earth Observation instruments such as MERIS (Medium-Spectral Resolution, Imaging Spectrometer) and AATSR (Advanced Along-Track Scanning Radiometer) from the European Space Agency's (ESA's) Envisat satellite platform have been commonly used for ocean color and sea surface temperature retrievals, respectively. Recent developments have enabled these instruments to now be applied to lakes to investigate lake water quality and lake surface water temperature (LSWT). These expansive spatiotemporal data sets simultaneously enable global assessment of environmental changes and present new statistical challenges.

GloboLakes (www.globolakes.ac.uk) is a 5-year Natural Environment Research Council consortium project involving 6 UK research groups. One of the aims of Globolakes is to investigate temporal coherence (similarities in major fluctuations in a set of time series) of water quality for 1000 lakes using a 20-year archive of satellite based spatial images (e.g. bi-monthly at 1° resolution). Determinands of interest include temperature, chlorophyll and coloured dissolved organic matter.

The aim of this paper is to present a dimensionality reduction and functional data approach for the clustering of remotely sensed spatiotemporal data from within lakes. The approach proposed is applied to LSWT data and provides a means of dealing with long periods of ice cover in some lakes.

1.1 Data

The ESA funded ARC-Lake project (MacCallum and Merchant, 2012) has employed the use of the AATSR instrument in order to derive observations of LSWT data for a large number of lakes across the globe. The data considered in this paper are spatially and temporally complete reconstructions of the ARC-Lake LSWT products for one lake, Lake Superior, from the ARC-Lake version 3 data-set (see www.geos.ed.ac.uk/arclake/data for details). The data-set for the lake is comprised of bi-monthly spatial images (each with 4094 pixels) for the 18-year period from 1995 to 2012. The time series includes periods of zeros which indicate time points with ice cover.

2 Methods

We propose initially to reduce the dimensionality of the individual lake spatiotemporal images. For each pixel within each lake, a functional data analysis (FDA) approach has been taken where each time series is represented as,

$$y_i(t) = G_i(t) + \epsilon_i(t) \quad i = 1, \dots n,\, t = 1, \dots, T$$

where G_i is a smooth curve and ϵ_i is a normally distributed independent random error term. The curve G_i is a spline function of degree d which can be expressed as a linear combination of B-splines, written in the following functional form for the spline

$$G_i(t) = \sum_{l=1}^{K+d-1} \beta_{i,l} B_l(t)$$

where $\beta_i = (\beta_{i,l}, \dots, \beta_{i,K+d-1})^T$ is a vector of real-valued coefficients, $B = (B_1(t), \dots, B_{k+d-1}(t))^T$ are the B-spline basis functions and K is the number of knots. To accommodate periods of ice cover we propose using an over-saturated basis to represent the curve, removing basis functions at periods of ice cover and then applying a smoothing parameter to fit the curve. Therefore, the β_i vector is estimated by least squares with a penalty:

$$(B^T B + \lambda D)^{-1} B^T \mathbf{y}$$

where λ is the smoothing parameter and D is a penalty matrix based on the integral of the squared second derivative of G. The curve G_i is then approximated by $\hat{G}_i(t)$.

Functional principal components analysis (FPCA) is then applied to the smooth curves, $\hat{G}$, to reduce dimensionality, and hence identify the dominant modes of variation in the data set. This provides a very computationally efficient way of exploring any underlying structure in the data, producing functional component scores:

$$f_{ik} = \int \xi_k(t) G_i(t) dt, \quad i = 1, \ldots, N,\ k = 1, \ldots, K,\ t = 1, \ldots, T$$

where ξ is eigenfuction k. Finally, a variety of clustering approaches have been applied (k-means, hierarchical and model-based (Fraley and Raftery, 1998)) to the functional component scores (formed as $f_{ik}w_j$). These have been adjusted to identify coherent regions within each lake, where w_j is a weight to account for the proportion of variability each functional component explains. Spatial correlation is accounted for via weights within the clustering procedure.

3 Results

Curves were fitted to the time series data for each pixel in Lake Superior using a cubic b-spline basis of 150 equally spaced functions. The smoothing parameter was set so that, after removal of basis functions in areas where there were zeros, there was one degree of freedom per 3 month season (66 in total). An example of a fitted curve for a single pixel is shown in the left panel of Figure 1 where points represent the data and the solid line represents the fitted smooth curve. As can be seen, there is good agreement between the fitted curve and the observations and the curve has captured the periods of ice cover well.

After applying FPCA and clustering, the statistically optimal number of clusters for describing the underlying variability in the pixel curves can be determined using approaches such as the L-curve or Gap statistic (Tibshirani *et al.*, 2001). Four clusters was found to be statistically optimal in terms of describing variability in Lake Superior. The clusters are displayed on a map of Lake Superior in Figure 1 (right). There were clear distinctions between the mean functions corresponding to these groups with the key discrepancy being the amplitude of the seasonal patterns each year. Group 1 pixel curves had the greatest amplitude (cooler in winter, warmer in summer) whilst Group 4 had the smallest.

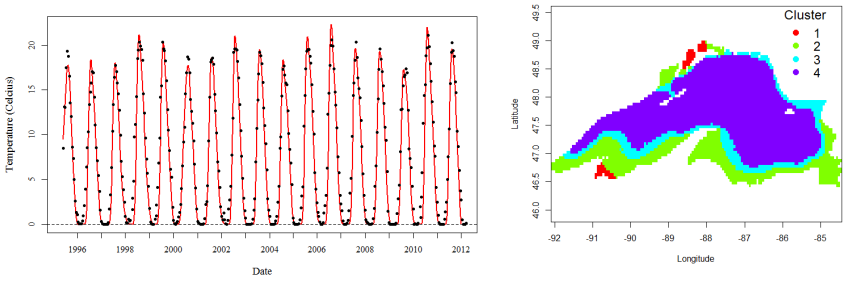

FIGURE 1. Left: Time series for a randomly selected pixel (points) and fitted spline function (solid line). Right: Map of Lake Superior, Canada, showing spatial distribution of estimated clusters.

4 Conclusions

The use of weighted functional PCs based on penalised over-saturated B-splines enables the data dimensions to be reduced substantially, while appropriately accounting for sequences of zeros in the time series. While methods have been developed for LSWT, we are currently applying and extending them for water quality measures such as chlorophyll.

5 Acknowledgements

O'Donnell, Scott and Miller were partly funded for this work through the NERC GloboLakes project (NE/J022810/1). The authors gratefully acknowledge the ARC lake project for access to the data.

References

Fraley, C. and Raftery, A. E. (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, **41** (8) 578 – 588.

MacCallum, S. N. and C. J. Merchant (2012). Surface water temperature observations of large lakes by optimal estimation. *Canadian Journal of Remote Sensing* **38** (1), 25 – 45.

R. Tibshirani, G. Walther and T. Hastie. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63** (2), 411 – 423.